DOCUMENT RESUME

ED 229 450 TM 830 360

AUTHOR Bogan, Evelyn Dood; Yen, Wendy M.
TITLE Detecting Multidimensionality and Examining Its Effects on Vertical Equating with the Three-Parameter Logistic Model.
PUB DATE Apr 83
NOTE 54p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Difficulty Level; *Equated Scores; Goodness of Fit; *Latent Trait Theory; *Multidimensional Scaling; *Simulation; Statistics; Testing; Test Items; Test Theory
IDENTIFIERS Chi Square Analysis; *Three Parameter Model; *Unidimensionality (Tests); Unidimensional Scaling; Vertical Equating

ABSTRACT

Four multidimensional data configurations and one unidimensional data configuration were simulated for three differences in mean difficulty between two tests to be equated. Two chi-square statistics, Q1 and Q2, were examined for their ability to detect multidimensionality. Results indicated that Q1 did not discriminate between any of the multidimensional configurations, nor did it discriminate any of the multidimensional conditions from any of the unidimensional conditions. The other chi-square statistic, Q2, overwhelmingly detected all three conditions of one of the four multidimensional configurations. Both horizontal and vertical equatings for the multidimensional configurations were as good as those for the unidimensional configurations when correlations were examined. When means were examined, the multidimensional tests usually had worse equatings than the unidimensional tests, particularly when the tests differed in difficulty. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED229450

Detecting Multidimensionality and Examining Its Effects on Vertical Equating with the Three-Parameter Logistic Model

Evelyn Doody Bogan
CTB/McGraw-Hill

Wendy M. Yen
CTB/McGraw-Hill

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

E.D. Bogan

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 830 360

Presented at the annual meeting of the American Educational Research Association, Montreal, April 1983.

Abstract

Four multidimensional data configurations and one unidimensional data configuration were simulated for three differences in mean difficulty between two tests to be equated. Two chi-square statistics, Q_1 and Q_2, were examined for their ability to detect multidimensionality. Results indicated that Q_1 did not discriminate between any of the multidimensional configurations, nor did it discriminate any of the multidimensional conditions from any of the unidimensional conditions. The other chi-square statistic, Q_2, overwhelmingly detected all three conditions of one of the four multidimensional configurations. Both horizontal and vertical equatings for the multidimensional configurations were as good as those for the unidimensional configuration when correlations were examined. When means were examined, the multidimensional tests usually had worse equatings than the unidimensional tests, particularly when the tests differed in difficulty.

Detecting Multidimensionality and Examining Its Effects on Vertical Equating with the Three-Parameter Logistic Model

In many testing situations, it is desirable or even necessary to compare test scores of different examinees or the same examinee across different forms or across levels of a test. It might be desired to compare scores across levels, to compare across multiple forms of one test, or to compare two entirely different tests. These different forms and especially different levels cannot be expected to be of equal difficulty and cannot be expected to be equally reliable. Therefore, scores on different forms and/or levels must be transformed in some way so that comparisons between test scores will be meaningful.

The Need for Vertical Equating

The need for vertical equating arises when scores from tests of unequal difficulty must be compared. Examples of such situations include following an individual's progress through one academic year, following an individual's progress across grades, comparing scores of different examinees taking different levels of the same test, or interpreting scores of students who are tested "out of level" because their ability levels are not the same as their grade placement.

Vertical equating is especially applicable to the case of a multilevel battery of tests whose levels are of increasing

difficulty. An advantage of such a series of tests graded in difficulty is that examinees ahead or behind their peers in their development can take the most appropriate level, increasing measurement accuracy. These out-of-level scores must then be scaled in some manner (vertically equated) if they are to serve a useful purpose.

The Unidimensionality Assumption

Despite the widespread use of vertical equating, any vertical equating results should be examined with caution since many problems remain unsolved. One problem that exists in many practical applications is the effect on vertical equating of violation of the unidimensionality assumption. This assumption, applied in this context, requires that each of the tests to be equated onto a common score scale must measure the same underlying (latent) trait or ability. In most equating models used to date, unidimensionality is assumed (i.e., the test items measure only a single latent trait or ability).

The assumption of unidimensionality frequently does not appear to be met, but the need for vertical equating exists. The issue then becomes that of the robustness of vertical equating to violation of the unidimensionality assumption.

Possible Causes of Violation of the Unidimensionality Assumption

Violation of the unidimensionality assumption could be caused by various factors. For example, the traits behind mathematical computation may change from level to level with mathematics performance possibly being differentially dependent upon school curriculum. Differencial probabilities of success could be due to lack of emphasis on or lack of introduction of some aspect of mathematical computation such as fractions or decimals (Loyd & Hoover, 1980). An achievement test in mathematics might require both reading skill and mathematical reasoning. An achievement test in Science might require both reading and knowledge of science facts.

Any factor that influences an examinee's score on a test, other than the one latent trait (ability) assumed for the one-dimensional model, will violate the assumption of unidimensionality. The existence of two or more cognitive traits influencing an examinee's response to an item is one such possible factor. Other possible factors are guessing, speededness, fatigue, cheating, deliberate random answering, or accidently overlooking or skipping an item.

Some of these factors can be controlled, reduced or eliminated, but the fact remains that vertical equating is needed and the unidimensionality assumption will be violated in many practical situations.

Test Analysis Model

The statistical model to be used in this study for analyzing

item responses is the three-parameter logistic model. This model assumes that an individual's performance on a test is influenced by only one important unobservable characteristic, θ, which is called a (latent) trait or ability. The three-parameter logistic model assumes that the probability of a correct response to item i by a person with ability level, θ, is:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta - b_i)}},$$

where a_i, b_i, and c_i are the discriminating power, difficulty, and lower asymptote or guessing parameter of item i, respectively.

The accuracy of item parameter estimation is affected by several things, including the accuracy of the estimation program (McKinley & Reckase, 1980), the size of the calibration sample (Hambleton, Swaminathan, Cook, Eignor & Gifford, 1978; Reckase, 1977), and the percent of test variance accounted for by the first factor found when the data are factor analyzed (Reckase, 1979). Violation of the unidimensionality assumption has been suggested as a problem in estimation of item parameters (Loyd & Hoover, 1980; Cook & Eignor, 1981). Therefore, if multidimensionality has an adverse effect on vertical equating when accurate parameter estimates are used and then this problem is further compounded by the effect of multidimensionality on estimating the item and person parameters of the three-parameter model, then the effect of multidmensionality on vertical equating could be quite extreme.

Method

The simulations begin with generation of two-dimensional data sets from a multidimensional model using prespecified parameters and end with an investigation into the effects on vertical equating of various multidimensional conditions. The main steps are:

1) Generate data.
2) Examine the data for multidimensionality.
3) Estimate parameters.
4) Perform equating.
5) Examine the accuracy of the equating.

Data Generation

The main question to be examined in this research is how robust three-parameter vertical equating is to violation of the unidimensionality assumption underlying the equating. The unidimensionality assumption is violated whenever the scores that are being equated are multidimensional in the sense that an examinee's score on a test is the result of more than one latent trait. The data can be the result of more than one latent trait and can also vary in the degree of the correlation that exists between these traits. Since infinitely many multidimensional data sets fulfilling these requirements are possible, this research project is necessarily limited to a few of the possibilities. Therefore, the number of dimensions and the choice of the degrees of correlation between traits were two of

the limitations placed on this research.

Degrees of Correlation. The choice of correlations was limited to that which seemed to be possible for published tests, Comprehensive Tests of Basic Skills, Forms U and V (CTBS/U, CTBS/V; CTB/McGraw-Hill, 1981). A correlation of .5 was chosen to simulate data sets on which Trait 1 and Trait 2 have medium correlation, while a correlation of .6 was used to simulate two traits that are more highly correlated. One unidimensional data set was also generated to be used as a criterion against which the multidimensional data sets could be compared.

Multivariate Model. Two-dimensional data sets were generated using a modified version of the multidimensional model described by Hattie (1982). This model is a three-parameter multivariate, logistic latent-trait model. It allows one to vary the levels of difficulty, guessing, discrimination, and the number of factors underlying the data. The multivariate logistic model is:

$$P_i(\underline{\theta}_j) = c_i + (1 - c_i)\left(1 + e^{-1.7\sum_{\ell=1}^{m} a_{i\ell}(\theta_{j\ell} - b_{i\ell})}\right)^{-1}$$

where $P_i(\underline{\theta}_j) = P_i(\theta_{j1}, \theta_{j2}, \ldots, \theta_{jm})$ and is the probability of a correct response to item i by a person j whose location in an m-dimensional latent space is described by abilities $\theta_{j1}, \theta_{j2}, \ldots, \theta_{jm}$.

$\theta_{j\ell}$ represents the ability of a person j on trait ℓ.

$a_{i\ell}$ is the discrimination of item i with respect to latent trait ℓ.

$b_{i\ell}$ is the difficulty of item i with respect to latent trait ℓ.

c_i is the guessing parameter for item i.

Since this research is concerned with two-dimensional data sets, m will be 2 with θ_{j1} representing the ability of person j on the first trait and θ_{j2} representing the ability of person j on the second trait. Note that when m = 1, this model reduces to the univariate logistic three-parameter model of Birnbaum (1968). This model was used with m = 1 to simulate the undimensional data. Thirty-item tests were used.

Difficulty Values. Base values of b_i were chosen to be reasonable for published tests (Lord, 1968; Ross, 1966; Yen, 1982). Lord (1968) reported values of b_i between -1.5 and 2.5 ($\overline{X}$ = .58, SD = .87); Ross (1966) had ranges of -3.14 to .49 and -2.07 to 1.42; and Yen (1982) reported a range of -1.46 to .95 ($\overline{X}$ = -.267, SD = .627). For this research, one set of b_i ($\overline{X}$ = -.028, SD = .818, with range -2.0 to +1.52) was chosen to have the following values: -2.00, -1.46, -1.25, -1.04, -.84, -.66, -.59, -.52, -.45, -.38, -.31, -.24, -.17, -.10, -.03, .04, .11, .18, .26, .32, .40, .49, .53, .62, .68, .74, .95, 1.02, 1.35, 1.52.

Three levels of differences in difficulty for the two tests to be equated were simulated. For the first set, both tests to be equated had the same difficulty values. This simulates a horizontal equating situation. The other two levels of difference in difficulty were chosen to be $\overline{b}_2 - \overline{b}_1 = 1.0$ and $\overline{b}_2 - \overline{b}_1 = 2.0$, where $\overline{b}_2$ is the mean difficulty of the harder test (Test 2) and $\overline{b}_1$ is the mean difficulty of the easier test (Test 1). Hereafter, the harder test will be called Test 2 and the easier test will be called Test 1. For $\overline{b}_2 - \overline{b}_1 = 1.0$, .5 was added to the base b_i

values to simulate the b_i for the harder test (Test 2) and .5 was subtracted from the base b_i values to simulate the b_i for the easier test (Test 1). Similarly, 1.0 was added to and subtracted from the base b_i values to create $\bar{b}_2 - \bar{b}_1 = 2.0$

Discrimination Values. Discrimination values in simulation studies are typically between 0.5 and 2.0 with a mean of 1.0. Warm (1978) states that typical values range from 0.5 to 2.5, usually between 1.0 and 2.0, and anything below .80 are not useful. Ree (1979) stated that items with discrimination greater than 2.0 are rarely found and any less than 0.5 do not discriminate well enough. He used values from 0.65 to 1.61 ($\bar{X}$ = 0.95, SD = 0.28). Hattie (1982) used simulated items with a range of 0.6 to 1.4 ($\bar{X}$ = 1.0, SD = 0.34). Lord (1968) found a range of 0.4 to 1.7($\bar{X}$ = 1.07, SD = 0.40) in a study of SAT Verbal items. For two multiple-choice tests, Ross (1966) used a range of 0.47 to 1.99 ($\bar{X}$ = 1.28) and .30 to 1.97($\bar{X}$ = 1.06), respectively.

For this study, a_i values for the simulated data sets were chosen to be between 0.5 and 2.0 ($\bar{X}$ = 1.03, SD = 0.3387): 0.5, 0.6, 0.6, 0.7, 0.7, 0.8, 0.8, 0.8, 0.8, 0.9, 0.9, 0.9, 0.9, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.1, 1.1, 1.1, 1.1, 1.2, 1.2, 1.2, 1.3, 1.4, 1.6, 1.8, 2.0.

Guessing Values. For a multiple-choice item with five alternatives, theoretically, the expected value of c is .20; however, values less than .20 are usually found in actual test items since some incorrect alternatives are deliberately made more attractive than others. Warm (1978) states that typical values range from 0.0 to 0.4 and recommends a range of 0.0 to 0.2.

Hattie (1982) used a range of 0.0 to 0.2, all 0.2, and all 0.0. Hambleton and Traub (1971) used a range of 0.0 to 0.2. Kingsbury and Weiss (1979) set all $c_i = 0.0$. Lord (1968) found a range of 0.04 to 0.2 ($\bar{X} = 0.16$, SD = 0.01) for his 80 SAT items. Most c_i values on CTBS/U (CTB/McGraw-Hill, 1981) are in the upper teens and lower twenties.

Therefore, in order to generate data that would be similar to the published tests to which the simulation results are to be applied, the following c_i values were chosen ($\bar{X} = 0.188$, SD = 0.021): 0.16, 0.16, 0.16, 0.16, 0.16, 0.16, 0.16, 0.17, 0.17, 0.17, 0.18, 0.18, 0.18, 0.19, 0.19, 0.19, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.20, 0.21, 0.21, 0.21, 0.22, 0.22, 0.22.

Item Pool. The base item pool consists of 30 items. Since the existence of two traits is being assumed, the model requires two discrimination parameters (a_{i1}, a_{i2}), two difficulty parameters (b_{i1}, b_{i2}), and one guessing parameter (c_i) per item.

For each test configuration, item parameters, a_{i1}, b_{i1}, a_{i2}, b_{i2}, and c_i, were randomly assigned to Traits 1 and 2 for both Tests 1 and 2, with the restriction that the desired correlations between parameters were approximated as closely as possible.

Test Configurations. Four two-trait data test configurations were chosen to be simulated. Each configuration was chosen as being a possible two-trait situation in a typical achievement test. As a starting point, CTBS/U and CTBS/V (CTB/McGraw-Hill, 1981) were examined from the point of view that the examinees' responses were the

result of two traits rather than one. Table 1 shows the desired trait and item parameter correlations for the simulated data sets. Table 2 shows possible real data situations where such correlations might exist.

Insert Tables 1 and 2 about here

For Configuration 1, both traits were assumed to be measured by both tests (Test 1, the easier test or level, and Test 2, the harder test or level). A high correlation ($r_{\theta_1\theta_2} = .6$) was assumed to exist between Trait 1 and Trait 2. All correlations for difficulty and discrimination, both within and between traits were assumed to be zero ($r_{a_1a_2} = 0$, $r_{a_1b_1} = 0$, $r_{a_2b_2} = 0$, $r_{b_1b_2} = 0$). This configuration was chosen as approximating a situation such as a Science test where an examinee's item responses could be the result of two traits: both reading ability and knowledge of science facts. Table 3 shows the item parameters used.

Insert Table 3 about here

For Configuration 2, both traits were assumed to be measured by both tests as in Configuration 1, except that both traits were not assumed to be measured by all items. Some items were assumed to be measuring only Trait 1 while having zero discrimination on Trait 2. Some items were assumed to measure Trait 2 only, having zero discrimination on Trait 1. Also, a medium correlation was assumed between traits ($r_{\theta_1\theta_2} = .5$), and discrimination was assumed to be negatively correlated across traits ($r_{a_1a_2} = -.8$).

This configuration was chosen to reflect a situation, such as

Social Studies, where the two traits might be reading ability and ability to understand graphs. A typical Social Studies test usually contains some items pertaining to graphs only. A few items may require both reading ability and an understanding of graphs. Other items will require reading only. The parameters for Configuration 2 are in Table 3.

For Configuration 3, both traits were again assumed to be measured by both tests and the correlation between traits was assumed to be .5. A few items were assumed to have low discrimination on Trait 1 and high discrimination on Trait 2. All other items were assumed to have a mixture of high, medium and low discrimination on Trait 1 and zero discrimination on Trait 2. This configuration was chosen as being similar to a situation such as Mathematics Computation, where Trait 2 is ability in computation problems involving decimals and Trait 1 is ability in computing all problems not involving decimals. A few items involving decimals were assumed to have low discrimination on Trait 1 and high discrimination on Trait 2 (the decimal trait). The 25 other items were assumed to not measure decimal ability so they have zero discrimination on Trait 2. See Table 3.

Configuration 4 was chosen as a situation in which the second trait is measured only by a few items on the harder test or level. The correlation between traits was assumed to be moderate ($r_{\theta_1\theta_2} = .5$). The correlation between difficulty and discrimination was assumed to be moderate ($r_{a_1b_1} = r_{a_2b_2} = .4$) for both tests on Trait 1. The correlation between difficulty

and discrimination for Trait 2 on the harder test was assumed to be high ($r_{a_2b_2}$ = .7). The difficulty parameters for the two traits were assumed to have a medium correlation ($r_{b_1b_2}$ = .5) on the harder test.

This configuration was chosen to reflect a situation, such as Mathematics Computation, where Trait 2 is ability with fractions (not measured by the easier test, measured by only a few items on the harder test) and Trait 1 is ability with other items (measured at both levels). See Table 3.

Configuration 5 is the unidimensional criterion. This configuration is simulated by setting m = 1 in the multidimensional model used to generate the data. Item parameters are shown in Table 3.

Data Conditions. Combining the above described configurations results in 15 data conditions: 5 (item parameter/correlation configurations) x 3 ($\bar{b}_2 - \bar{b}_1$) values.

Simulees. The simulated data sets include three groups of simulees: 2,000 simulees of low ability, 2,000 of middle ability, and 2,000 of high ability. The 2,000 theta values for each of the three levels were generated using the IMSL multivariate normal random deviate generator GGNSM (IMSL, 1979). In each case a normal distribution is assumed ($\bar{\theta}$ = -0.57, SD_θ = 1.0 for the low ability group, $\bar{\theta}$ = 0.0, SD_θ = 1.0 for the middle ability group and $\bar{\theta}$ = 0.57, SD_θ = 1.0 for the high ability group). These differences in means were chosen to be similar

to the differences between ability levels in published tests (CTBS/U Levels E & F, and H & J). A total group consisting of these 6,000 simulees pooled was also formed. For the 15 data conditions, response vectors were generated for each of the three groups of simulees for tests of 30 items each.

Anchor Test. Recall that for the vertical equating situations, Test 1 is the easier test and Test 2 is the harder test of the pair of tests to be equated. For the horizontal equating conditions, Test 1 is either test and Test 2 is the other. For each of the 15 data conditions, a 30 item anchor test was chosen based on the p-values. Each anchor test was composed of the 15 items of Test 1 with the lowest p-values and the 15 items of Test 2 with the highest p-values.

Response Vectors. Using the prespecified "true" parameters (a_{i1}, a_{i2}, b_{i1}, b_{i2}, c_i, θ_{j1}, θ_{j2}) and the multidimensional model, $P_i(\underline{\theta}_j)$ was computed for each simulee. From these $P_i(\underline{\theta}_j)$ values, (0, 1) responses, u_{ijk} were generated for each item i, simulee j, and test k where u_{ijk} is 1 if a random number is less than or equal to $P_i(\underline{\theta}_j)$ or 0 otherwise. The random number was generated from a uniform distribution using IMSL subroutine GGUBS (IMSL, 1979).

Responses were generated for all simulees for both tests (Test 1 and Test 2). For the medium ability group, responses to the anchor test were separated from Test 1 and Test 2 to be used for parameter estimation and equating. Responses for all simulees to both tests were used to examine the accuracy of the equating.

Data Verification. The means and standard deviations of the number-correct scores, item difficulties (p-values), and the KR-20s from the simulated tests were compared to statistics for real tests, Comprehensive Tests of Basic Skills, Form U (CTB/McGraw-Hill, 1981), in order to verify that the simulations are realistic.

Detecting Multidimensionality

For each set of generated data, the response of the following fit measures to multidimensionality was examined: Q_1 (Yen, 1981), and Q_2, a statistic similar to the van den Wollenberg (1982) Q_2 statistic. The Q_2 used here has been generalized to apply to the three-parameter model.

For the Q_1 statistic, the examinees are ordered according to their ability estimates. The range of ability estimates is then divided into ten approximately equal categories and the statistic is computed by:

$$Q_1 = \sum_{k=1}^{10} \frac{N_k(O_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})}$$

where N_k is the number of examinees in cell k,

O_{ik} is the observed proportion of examinees in cell k that passes item i, and

E_{ik} is the expected/predicted proportion of examinees in cell k that passes item i.

$$E_{ik} = \frac{1}{N_k} \sum_{j \in k}^{N_k} \hat{P}_i(\hat{\theta}_j),$$

where $\hat{P}_i(\hat{\theta}_j)$ is evaluated using the trait estimate for examinee j, and the item parameters estimated for item i. The summation is over examinees in cell k.

When the model being examined is true, Q_1 should be distributed approximately as a chi-square variable with 10-m degrees of freedom, where m = 3, the number of item parameters estimated by the model (Yen, 1981).

The second fit statistic used is Q_2, a statistic similar to the van den Wollenberg (1982) Q_2 statistic. This statistic is a chi-square statistic developed by Yen (Note 1) generalized to apply to the three-parameter model.

Q_2 is a test of local independence which is based on the observed numbers in two-way contingency tables of each pair of items for each overall test score group (cell).

		item j: 0	item j: 1	Total
item i	0	$N_{\bar{i}\bar{j}r}$	$N_{\bar{i}jr}$	$N_{\bar{i}r}$
	1	$N_{i\bar{j}r}$	N_{ijr}	N_{ir}
	Total	$N_{\bar{j}r}$	N_{jr}	N_{r}

where N_r is the number of examinees in cell r,

$N_{\bar{i}\bar{j}r}$ is the number of examinees in this cell who failed both items,

$N_{\bar{i}jr}$ is the number of examinees in this cell who failed item i and passed item j, etc.

The fit statistic for the r-th cell is:

$$Q_{2r} = \frac{\left[N_{\bar{i}\bar{j}r} - E(N_{\bar{i}\bar{j}r})\right]^2}{E(N_{\bar{i}\bar{j}r})} + \frac{\left[N_{\bar{i}jr} - E(N_{\bar{i}jr})\right]^2}{E(N_{\bar{i}jr})} + \frac{\left[N_{i\bar{j}r} - E(N_{i\bar{j}r})\right]^2}{E(N_{i\bar{j}r})} + \frac{\left[N_{ijr} - E(N_{ijr})\right]^2}{E(N_{ijr})}$$

$$Q_2 = \sum_r Q_{2r}$$

The expected values are obtained using only the observed numbers in the two-way contingency table (a traditional chi-square for association in a two-by-two contingency table).

$$E(N_{ijr}) = N_{ir}N_{jr}/N_r$$

$$E(N_{\bar{i}jr}) = N_{\bar{i}r}N_{jr}/N_r$$

$$E(N_{i\bar{j}r}) = N_{ir}N_{\bar{j}r}/N_r$$

$$E(N_{\bar{i}\bar{j}r}) = N_{\bar{i}j}N_{\bar{j}r}/N_r$$

Note that Q_{2r} equals N_r times the squared phi-coefficient between item scores in cell r.

Although Yen (Note 1) found some evidence that Q_2 detected multidimensionality for small groups of items that appeared to differ in content from the rest of the test, the overall mean Q_2 for the unidimensional data sets was nearly as large as the overall mean Q_2 for the multidimensional data sets. The percent of Q_2 values greater than 23.2 was much more useful in detecting multidimensional data sets. Results from simulations showed that Q_2 may approximate a chi-square distribution with ten degrees of freedom when the data are unidimensional. For a chi-square statistic with ten degrees of freedom, one percent of its values are greater than 23.2.

Criterion for Selecting a Fit Measure. The criterion used for determining which fit measures are capable of detecting multidimensionality is the size of the fit measure. As the correlation between the traits decreases the fit measure should increase. In other words, as the data become more multidimensional, the value of the fit measure should be larger.

Parameter Estimation

For each of the 15 data conditions, item parameters were estimated in two separate runs with LOGIST (Wingersky, Barton, & Lord, 1982). Table 4 shows the LOGIST runs used for

Insert Table 4 about here

parameter estimation. Item parameters for Test 1 and the anchor were estimated using the responses from the low and medium ability groups combined. The 30 items from Test 2 were considered as "not reached" (NR) items for the low ability group. The 15 test 2 and the 15 Test 1 items not in the anchor were considered as NR items for the medium ability group. Similarly, item parameters for Test 2 and the anchor were estimated using the responses from the medium and high ability groups combined. Items from Test 1 were considered as NR for the high ability group. The 30 Test 2 and Test 1 items not in the anchor were NR for the medium ability group.

This allows combined samples of 4,000 examinees, a sample size that has been found to be adequate for obtaining very stable item parameter estimates (Yen, in press, p. 19).

Since separate pairs of LOGIST runs were made for each of the 15 data conditions, the result was two sets of estimated item parameters and estimated thetas per pair of tests. For each of these 15 conditions, equating was performed and the accuracy of the equating was examined.

Equating Between Levels

A second set of LOGIST runs using fixed item parameters was made in order to examine the between level equating. Since the anchor-test design was used for equating, Test 1 was taken by one group (low ability), Test 2 was taken by another group (high ability), and the anchor was taken by a third group (medium ability). In the first LOGIST run the parameters for Test 1 and the anchor were estimated using the responses from the low ability and the medium ability groups. Test 1 and the anchor test have items in common (the 15 hardest items of Test 1) and the parameters were estimated in one LOGIST run, leaving the parameters for Test 1 and the anchor test on the same scale (i.e., equated). In Run 2, the parameters for Test 2 and the anchor were estimated using the responses from the high and medium ability groups. Since Test 2 and the anchor have common items (the 15 easiest items of Test 2) and only one LOGIST run was used, the parameters for Test 2 and the anchor are on the same scale (i.e., equated). The means and standard deviations of the $\hat{b}_i$ values for the anchor items were equated to complete the interlevel linking:

a. $M_{\hat{b}_1}$, $SD_{\hat{b}_1}$, $M_{\hat{b}_2}$ and $SD_{\hat{b}_2}$ were calculated where $M_{\hat{b}_1}$ and $SD_{\hat{b}_1}$ refer to the mean and standard deviation of the item difficulty parameter, $\hat{b}_i$, for the anchor items as estimated in Run 1 and $M_{\hat{b}_2}$ and $SD_{\hat{b}_2}$ refer to the mean and standard

deviation of the $\hat{b}_i$ for the anchor items as estimated in Run 2.

b. Let $K_1 = SD_{\hat{b}_1} / SD_{\hat{b}_2}$

and

$$K_2 = M_{\hat{b}_1} - K_1 M_{\hat{b}_2}$$

c. The estimated parameters for all a, b, c, and θ on the items from Run 2 were then adjusted as follows:

$$\hat{c}_2^* = \hat{c}_2$$

$$\hat{a}_2^* = \hat{a}_2 / K_1$$

$$\hat{\theta}_2^* = \hat{\theta}_2 K_1 + K_2$$

$$\hat{b}_2^* = \hat{b}_2 K_1 + K_2$$

The estimated parameters for Run 2 have then been placed on the same scale as those for Run 1 (i.e., Test 2 is equated to Test 1).

Between Level Comparisons.

For each of the 15 data conditions, thetas (θ) were estimated in two separate runs with parameters (a, b, c) fixed.

Insert Table 5 about here

Table 5 shows the LOGIST runs used. In Run 1, thetas were estimated for Test 1 using item responses for all three ability

levels on Test 1 and NR for Test 2. Item parameters were fixed at the values estimated in the parameter estimation LOGIST runs. Similarily, in Run 2, thetas were estimated for Test 2, using item responses on Test 2, NR on Test 1, and fixed item parameters.

Since the equated fixed item parameters were used, Test 1 and Test 2 are equated after these LOGIST runs. The accuracy of the equating must then be examined.

Examining the Accuracy of the Equating

The main purpose of this research is to examine the robustness of three-parameter vertical equating to violation of the unidimensionality assumption underlying the equating. It is hypothesized that the accuracy of vertical equating is affected by data that are multidimensional. The estimation of the item parameters (a, b, c) is also affected by violation of the unidimensionality assumption. The data used in this research are generated by a multidimensional model. Several situations are simulated in which simulees' responses to items are influenced by more than one trait. Three levels of differences between mean difficulty on the two tests to be equated were simulated. The accuracy of the parameter estimation was examined in light of these two effects: the effect of each individual multidimensional data condition, and the size of the difference between mean difficulty on the two tests that were equated.

Accuracy of the Equating. The accuracy of the equatings was examined by comparing the estimated thetas for all simulees on the easier test with the estimated thetas for all simulees on the harder test in terms of standardized root mean squared differences (SRMSD), standardized differences between means (SDM), the ratio of the standard deviations, correlations, and tau equavalence.

The RMSD between pairs of estimated thetas, scaled by the variances of the two trait estimates being compared, was computed as follows:

$$\text{SRMSD} = \sqrt{\frac{\frac{1}{n}\sum_{j=1}^{n}(\hat{\theta}_{1j} - \hat{\theta}_{2j})^2}{(S_1^2 + S_2^2)/2}}$$

where $\hat{\theta}_{1j}$ and $\hat{\theta}_{2j}$ are the estimates of the trait for person j (j = 1, ..., n simulees) for Tests 1 and 2 respectively,

and S_1^2 and S_2^2 are the variances of the theta estimates for Test 1 and 2 respectively.

The standardized difference between means (SDM), an estimate of overall bias between the estimated thetas, is the difference in mean thetas for Test 1 and Test 2 divided by a pooled estimate of the standard deviations:

$$\text{SDM} = \frac{\bar{\hat{\theta}}_1 - \bar{\hat{\theta}}_2}{\sqrt{\left(S_1^2 + S_2^2\right)/2}}$$

where $\bar{\hat{\theta}}_1$ is the mean of the estimated thetas on Test 1,

$\bar{\hat{\theta}}_2$ is the mean of the estimated thetas on Test 2,

S_1^2 is the variance of the estimated thetas on Test 1,

and S_2^2 is the variance of the estimated thetas on Test 2.

A third method of checking the accuracy of the equatings was to examine the tau-equivalence of the equated estimated thetas (Yen, in press). Two scores, X_1 and X_2 are tau-equivalent if

$$\mathcal{E}(X_1|T) = \mathcal{E}(X_2|T),$$

where T is the true score. The local equating bias of $\hat{\theta}_2$ and $\hat{\theta}_1$ is defined as

$$B(\hat{\theta}_1, \hat{\theta}_2|T) = \mathcal{E}(\hat{\theta}_1|T) - \mathcal{E}(\hat{\theta}_2|T).$$

To estimate local biases, simulees were rank ordered on the basis of the average of the two trait values and grouped into five cells. The estimate of local equating bias is

$$\hat{B}_m(\hat{\theta}_1, \hat{\theta}_2|T) = \frac{\bar{\hat{\theta}}_{1m} - \bar{\hat{\theta}}_{2m}}{\sqrt{(S_1^2 + S_2^2)/2}} \quad \text{for cell m}$$

where $\bar{\hat{\theta}}_{1m}$ and $\bar{\hat{\theta}}_{2m}$ are the means of $\hat{\theta}_1$ and $\hat{\theta}_2$ for cell m.

The standardized difference between means was found for each of the five cells. The average of the absolute values of the standardized mean differences for the five cells was used as a summary measure of local bias.

Results

True Item Parameters. Table 6 shows the desired and attained correlations among the item parameters used for generating the two-trait data. All attained correlations are within ±.09 of the desired

Insert Table 6 about here

correlations, hence the attained correlations appear to be acceptable.

Simulated Thetas. Tables 7 and 8 contain the desired and attained correlations, means, and standard deviations for the true (generated) thetas. The attained correlations between generated

Insert Tables 7 & 8 about here

thetas on the two traits are all within .03 of the desired correlations. Mean thetas are all within .06 of the desired mean ability for all three ability levels and the standard deviations of the generated true thetas are all within .04 of the desired standard deviations. The simulated thetas appear to be acceptable.

Simulated Item Responses. Table 9 contains the means and standard deviations of the number-correct scores and the KR-20

values for the pairs of simulated tests. The simulated tests

Insert Table 9 about here

appear to be realistic, although for the $\bar{b}_2 - \bar{b}_1 = 2$ conditions, Tests 1 and 2 differ a great deal in difficulty. Recall that Test 1 and Test 2 are simulated to have equal difficulty when $\bar{b}_2 - \bar{b}_1 = 0$. When $\bar{b}_2 - \bar{b}_1 = 1$ or $\bar{b}_2 - \bar{b}_1 = 2$, Test 2 is the harder of the pair of tests. Test 1 and Test 2 do have very similar difficulties for all configurations when $\bar{b}_2 - \bar{b}_1 = 0$. For all configurations, Test 1 appears easiest and Test 2 appears hardest when $\bar{b}_2 - \bar{b}_1 = 2$.

Tables 10 and 11 contain the mean chi-squares for the 30 data conditions. Table 10 contains the mean Q_1 chi-squares. There is

Insert Tables 10 & 11 about here

no evidence that Q_1 detects the multidimensionality of the original data sets. Table 11 contains the mean Q_2 (multidimensional) chi-squares and the percent of the Q_2 values that are greater than 23.2. The Q_2 statistic clearly detects the multidimensionality of all conditions of Configuration 2. This configuration has mean Q_2 values ranging from 18.3 to 25.8 compared to ranges from 9.1 to 15.8 for the other three multidimensional configurations and the unidimensional configuration. Configuration 2 also has 20-47 percent of Q_2 values greater than 23.2 compared to 4-14 percent for the other configurations.

Table 12 contains the means, standard deviations, and

Insert Table 12 about here

correlations of the estimated thetas for Test 1 versus Test 2. The correlations between the easier and the harder tests are mostly higher than .80. The poorest equatings appear to be for the conditions when $\bar{b}_2 - \bar{b}_1 = 2$. In these conditions, the correlations range from .70 to .75. The correlations from the multidimensional data sets are as good as the unidimensional correlations.

The standardized root mean squared differences (SRMSD), standardized differences between the means (SDM), ratios of standard deviations, and the summary measures of local bias are presented in Table 13. The SRMSD follow the same pattern as the

Insert Table 13 about here

correlations. The SRMSD are mostly around .60, except for the conditions when the mean difference in difficulty between the two tests is greatest. In these conditions, the most extreme vertical equating situation examined here, the SRMSD ranges from .72 to .79.

The SDM values are small for the unidimensional conditions. In multidimensional Configurations 1 and 2, the SDM values are larger than those for the unidimensional configuration, particularly for the vertical equatings. For Configurations 3 and 4, the SDM values are small, except for the most extreme vertical equating situation ($\bar{b}_2 - \bar{b}_1 = 2$). Test 2 means usually are higher than Test 1 means.

The local bias measure follows the trends for the SDM values.

Local equating bias by quintile is shown in Table 14. The con-

Insert Table 14 about here

ditions with large SDM values and large overall amounts of local bias in Table 13 tended to have large amounts of local bias in all quintiles. In the extreme vertical equating conditions, a moderate to high amount of equating bias appears in all five quintiles for all of the multidimensional conditions (except Configuration 1, for which there is no data). This is consistent with the correlations, which indicated that the poorest equating occurred in the most extreme vertical equating situations.

In item response theory (IRT), the standard error of measurement is the square root of the reciprocal of the information function. One of the advantages of IRT is that it allows different standard errors of measurement for every theta, since the model predicts the standard error of measurement as a function of trait value and item parameter values. Figures 1 and 2 show the predicted and

Insert Figures 1 & 2 about here

the observed standardized standard errors of measurement by quintile. Figure 1 is typical of the horizontal equating conditions and of the less extreme vertical equating conditions. Figure 2 is typical of the most extreme vertical equating conditions, except for Configuration 3 and the unidimensional criterion. For these two conditions (Configuration 3 and the unidimensional criterion), the observed SSE in the lowest quintile is around .7, while the predicted is larger than 5.6. The model accurately predicts the SSE values for all but the lowest quintile. The reason for the poor prediction in the lowest quintile of these two conditions is not apparent.

Discussion

The purpose of this research is to look at some measures for detecting multidimensionality and to examine the effects of various multidimensional data configurations on vertical equating with the three-parameter model. Four two-trait data configurations and one unidimensional criterion were chosen. For each of these five configurations, one horizontal ($\bar{b}_2 - \bar{b}_1 = 0$) and two vertical ($\bar{b}_2 - \bar{b}_1 = 1$, $\bar{b}_2 - \bar{b}_1 = 2$) equating conditions were simulated. The simulated item parameters and thetas were well within acceptable limits of the desired values. Means, standard deviations, and KR-20 values of number-correct scores indicated that all the conditions simulated realistic test configurations, although the $\bar{b}_2 - \bar{b}_1 = 2$ condition produced tests that had quite different difficulties.

Two fit measures were examined for their ability to detect multidimensional data conditions. The criterion for determining how well a fit measure could detect multidimensionality was the size of the fit measure; the value of the fit measure should increase as the data become more multidimensional. Two degrees of correlation between traits were simulated: .50 and .60.

It was found that Q_1 did not discriminate between the multidimensional conditions and the unidimensional criterion, nor did it discriminate between any of the multidimensional conditions. This supports Yen's findings (Note 1) that Q_1 has the same magnitude no matter how multidimensional the data are. This also supports van den Wollenberg's (1982) findings that many of the chi-square statistics used to verify unidimensionality must be questioned.

Q_2 also did not satisfy the criterion for determining how well a fit measure detects multidimensionality. (It did not increase as the degree of correlation between the two traits decreased.) However, the difference in trait correlations was not large. Q_2 did overwhelmingly discriminate between Configuration 2 (for all conditions) and the other configurations, including the unidimensional criterion. Mean Q_2 for Configuration 2 ranged from 18.3 to 25.8 compared to 9.1 to 15.8 for all other conditions. The percent of Q_2 greater than 23.2 was 20-47 compared to 4-14 percent for all other conditions.

By examining the item discriminations in Table 3 it is seen that Configuration 2 has the strongest second factor; roughly half of the items "load" (discriminate) on one factor and half on the other factor. It would be useful to factor analyze the item intercorrelations to corroborate this apparently strong two factor structure.

The fact that the Q_2 values were so large for Configuration 2 supports both Yen's (Note 1) findings that Q_2 is sensitive to multidimensionality when the three-parameter model is used and van den Wollenberg's (1982) findings that his Q_2 for the Rasch model successfully detected multidimensionality.

Yen (Note 1) found evidence that Q_2 could detect multidimensionality for small groups of items that were suspected a priori to have different content from the rest of the test even when the overall mean Q_2 did not indicate the existence of multidimensionality. All five configurations (30 conditions) used in this research are being examined from this point of view. Items are being grouped as having similar content (discriminations) according to the specific guidelines that were used to develop each data configuration. Values of Q_2 for within and between "content" groups will be compared.

Despite the fact that Configuration 2 had the clearest two-dimensional structure, it did not have outstandingly bad vertical equating results. This effect may have occurred because both dimensions were measured in both test levels. Configuration 4 had a few items that discriminated on the second factor only on the harder test, but the vertical equating results for this configuration were relatively good. Perhaps if a configuration were created with a larger proportion of items discriminating on Dimension 1 in Test 1 and on Dimension 2 in Test 2, vertical equating results would deteriorate more markedly.

In general both the horizontal and vertical equatings for the multidimensional configurations were as good as those for the unidimensional configuration when correlations or SRMSD values were examined. When means were examined using SDM or local bias statistics, the multidimensional tests usually had worse equatings than the unidimensional tests, particularly when the tests differed in difficulty.

Reference Note

1. Yen, W.M. A Fit Measure That Detects Multidimensionality. Unpublished manuscript, 1982 (Available from the author, CTB/McGraw-Hill, 2500 Garden Road, Monterey, CA 93940).

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M., & Novick, M. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968. Chapters 17-20.

Comprehensive Tests of Basic Skills, Forms U & V. Monterey, CA: CTB/McGraw-Hill, 1981.

Cook, L.L., & Eignor, D.R. Score equating and item response theory: Some practical considerations. Paper presented at the annual meeting of AERA - NCME, Los Angeles, 1981.

Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R., & Gifford, J.A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Hattie, J. Decision criteria for determining unidimensionality: An empirical study. Armidale, N.S.W., Australia: Center for Behavioral Studies, 1982.

IMSL Library (7th Ed.). Houston, TX: International Mathematical & Statistical Libraries, 1979.

Kingsbury, G.G., & Weiss, D.J. Relationships among achievement level estimates from three item characteristic curve scoring methods. (Research Report 79-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, April 1979.

Lord, F.M. Analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Menlo Park, CA: Addison-Wesley, 1968.

Loyd, B.H. & Hoover, H.D. Vertical equating using the Rasch model. Journal of Educational Measurement, 1980, 17, 179-193.

McKinley, R.L. & Reckase, M.D. A comparison of the ANCILLES and LOGIST parameter estimation procedures for the three-parameter logistic model using goodness of fit as a criterion. (Research Report 80-2). Columbia: University of Missouri, Department of Educational Psychology, December, 1980.

Reckase, M.D. Ability estimation and item calibration using the one and three parameter logistic models: A comparative study (Research Report 77-1). Columbia, MO: University of Missouri, Department of Educational Psychology, 1977.

Reckase, M.D. Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 1979, 4, 207-230.

Ree, M.J. Measurement and latent-trait theory. Paper presented at the 1979 AFHRL Conference on Human Assessment, March 1979.

Ross, J. An empirical study of a logistic mental test model. Psychometrika, 1966, 31, 325-340.

van den Wollenberg, A.L. Two new test statistics for the Rasch model. Psychometrika, 1982, 47, 123-140.

Warm, T.A. A primer of item response theory. Oklahoma City, OK: U.S. Coast Guard Institute, 1978.

Wingersky, M.A., Barton, M.S., & Lord, F.M. LOGIST User's Guide. Princeton, N.J.: Educational Testing Service, 1982.

Yen, W.M. Using Simulation Results to Choose a Latent Trait Model. Applied Psychological Measurement, 1981, 5, 245-262.

Yen, W.M. Use of three-parameter item response theory in the development of CTBS, Form U, and TCS. Paper presented at the National Council on Measurement in Education, New York, March 1982.

Yen, W.M. Tan equivalence and equipercentile equating. Psychometrika, in press.

TABLE 1

Correlations for Simulated Data Sets

Configuration	Test 1		Test 2	
	Trait 1	Trait 2	Trait 1	Trait 2
1. Both traits are measured by both tests.	$r_{\theta_1\theta_2} = .6$			
High correlation between traits.	$r_{a_1b_1} = 0$	$r_{a_2b_2} = 0$	same as lower level	
Randomly assign a, b, c, to items.	$r_{a_1a_2} = 0$	$r_{b_1b_2} = 0$		
2. Both traits are measured by both tests.	$r_{\theta_1\theta_2} = .5$			
Medium correlation between traits.	$r_{a_1b_1} = 0$	$r_{a_2b_2} = 0$	same as lower level	
Discrimination negatively correlated across traits.	$r_{a_1a_2} = -.8$	$r_{b_1b_2} = 0$		
3. Both traits measured by both tests.	$r_{\theta_1\theta_2} = .5$			
A few items low discrimination on Trait 1, high discrimination on Trait 2.	$r_{a_1b_1} = -.3$	$r_{a_2b_2} = 0$	same as lower level	
All other items have hi/med/lo discrimination on Trait 1, 0 discrimination on Trait 2.	$r_{a_1a_2} = -.5$	$r_{b_1b_2} = 0$		
4. Second trait measured only by harder test by a few items.	$r_{\theta_1\theta_2} = .5$		$r_{\theta_1\theta_2} = .5$	
Correlation between difficulty and discrimination on Trait 1 is medium, on Trait 2 is high.	$r_{a_1b_1} = .4$	$r_{a_2b_2} = .5$	$r_{a_1b_1} = .4$	$r_{a_2b_2} = .7$
Medium correlation for difficulty across traits (on harder test).	$r_{a_1a_2} = 0$	$r_{b_1b_2} = 0$	$r_{a_1a_2} = 0$	$r_{b_1b_2} = .5$
5. Undimensional Criterion				

TABLE 2

Applications to Real Data

Test	Test 1		Test 2	
	Trait 1	Trait 2	Trait 1	Trait 2
1. Science	reading	science facts	reading	science facts
Language Mechanics	reading	punctuation	reading	punctuation
2. Social Studies	reading	graphs	reading	graphs
Language Mechanics	punctuation	capitals	punctuation	capitals
3. Mathematics Computation	other items	decimals	other items	decimals
4. Mathematics Computation	other items	fractions	other items	fractions

TABLE 3

CONFIGURATION 1 TWO TRAITS
BASE ITEM PARAMETERS

EASY TEST					HARD TEST				
AL1	AL2	BL1	BL2	CL	AH1	AH2	BH1	BH2	CH
0.80	0.80	0.68	-1.25	0.18	0.90	1.00	0.62	-2.00	0.16
2.00	0.80	0.18	1.52	0.20	0.90	1.00	-0.03	0.04	0.18
1.00	2.00	0.62	-0.38	0.18	2.00	1.40	-0.31	-0.52	0.17
1.10	1.00	-0.17	-2.00	0.20	1.00	1.10	1.35	0.49	0.20
0.50	0.90	0.32	-0.03	0.16	0.60	1.80	1.02	0.18	0.19
1.20	1.80	0.40	0.74	0.20	1.00	2.00	0.95	-1.25	0.20
1.00	1.00	0.11	0.18	0.16	0.80	1.10	-0.84	0.53	0.16
1.20	0.50	-0.66	0.62	0.16	1.00	0.50	-2.00	-0.10	0.18
1.20	1.20	-0.31	0.11	0.18	1.10	0.90	0.26	-0.31	0.20
0.90	0.80	-0.84	0.32	0.20	0.60	1.10	-0.10	0.32	0.22
1.10	1.20	-2.00	-0.84	0.19	1.10	0.60	0.32	0.11	0.22
0.60	0.70	-0.45	-1.46	0.19	1.00	1.00	-0.45	-0.84	0.20
0.80	0.60	0.49	-0.17	0.16	0.50	1.00	0.49	-0.24	0.21
1.00	1.10	-0.59	0.68	0.22	0.70	0.70	0.68	0.40	0.16
0.90	1.10	-0.52	0.40	0.17	1.00	1.00	-1.46	1.02	0.17
0.80	1.10	-1.25	-0.59	0.20	0.80	1.60	-0.59	-1.04	0.19
0.90	1.00	-0.24	-1.04	0.20	1.20	0.80	-0.52	-1.46	0.16
0.70	0.80	1.52	1.35	0.20	1.60	0.90	0.53	-0.17	0.20
0.70	1.20	-0.10	1.02	0.16	0.90	0.80	-0.17	-0.66	0.19
1.10	0.60	-0.38	0.26	0.16	1.80	0.90	0.04	-0.03	0.17
0.60	1.40	-1.04	0.49	0.21	0.70	1.20	-1.04	0.62	0.16
1.60	0.90	0.95	-0.24	0.21	1.30	0.80	-1.25	-0.45	0.18
1.80	1.10	-1.46	0.95	0.20	0.80	0.90	-0.24	0.95	0.16
1.00	1.30	1.02	0.53	0.22	1.20	1.20	0.18	1.35	0.20
0.90	0.90	0.53	-0.66	0.22	0.90	1.20	0.74	0.26	0.21
1.30	1.60	1.35	-0.52	0.21	1.10	0.80	-0.66	-0.59	0.16
1.40	0.90	0.04	-0.10	0.16	0.80	0.60	-0.38	-0.38	0.20
0.80	1.00	-0.03	0.04	0.17	1.10	0.70	0.11	0.74	0.21
1.00	0.70	0.26	-0.31	0.17	1.40	1.30	0.40	1.52	0.20
1.10	1.00	0.74	-0.45	0.19	1.20	1.10	1.52	0.68	0.22

Note: AL1 & BL1 are the discrimination and difficulty on the easy test for Trait 1.

AL2 & BL2 are the discrimination and difficulty on the easy test for Trait 2.

AH1 & BH1 are the discrimination and difficulty on the hard test for Trait 1.

AH2 & BH2 are the discrimination and difficulty on the hard test for Trait 2.

TABLE 3 (CONT.)

CONFIGURATION 2 TWO TRAITS
BASE ITEM PARAMETERS

EASY TEST					HARD TEST				
AL1	AL2	BL1	BL2	CL	AH1	AH2	BH1	BH2	CH
0.00	1.10	-0.66	0.04	0.17	0.00	1.20	-1.46	-0.66	0.18
0.00	1.20	0.95	0.68	0.21	0.60	1.30	0.62	0.49	0.20
0.60	1.80	-0.45	-0.31	0.17	0.60	0.00	0.40	-0.10	0.17
0.70	0.00	-1.46	0.40	0.18	0.70	0.00	0.95	-0.45	0.20
0.00	1.30	0.11	0.32	0.20	0.00	1.80	-2.00	1.02	0.16
0.80	0.00	-0.31	1.02	0.16	0.80	0.00	1.52	1.52	0.20
0.80	0.00	-0.24	1.52	0.19	0.80	0.00	0.04	-0.59	0.22
0.80	1.10	0.40	-0.45	0.19	0.80	0.00	0.74	-1.25	0.21
0.00	1.40	0.74	0.26	0.20	0.80	1.10	0.68	-0.31	0.16
0.90	0.00	0.18	0.53	0.20	0.00	1.00	-0.24	-0.24	0.20
0.90	0.90	-0.17	-0.84	0.20	0.90	0.00	0.18	0.68	0.18
0.90	0.00	-2.00	0.11	0.19	0.00	1.10	1.35	1.35	0.20
0.00	1.00	1.35	0.95	0.17	0.00	1.20	-1.04	-0.52	0.16
1.00	0.00	0.49	-1.25	0.22	1.00	0.00	0.49	0.04	0.17
0.00	1.20	0.32	-0.17	0.21	0.00	1.10	-0.66	0.18	0.16
1.00	0.00	-0.84	-0.10	0.16	1.00	0.00	0.32	0.62	0.19
0.00	1.00	-0.52	0.74	0.16	0.00	1.40	-0.84	0.32	0.20
0.00	1.00	-0.10	-0.59	0.20	1.00	0.00	-0.45	0.11	0.19
1.10	0.00	-0.59	0.18	0.20	1.10	0.90	0.26	-1.04	0.22
1.10	0.00	-1.04	-0.66	0.16	1.10	0.00	-0.03	0.95	0.17
0.00	0.80	0.04	-0.38	0.18	0.00	0.90	0.11	-0.03	0.16
1.10	0.50	0.62	0.49	0.22	0.00	0.90	0.53	0.53	0.16
0.00	0.80	1.02	-0.52	0.16	1.20	0.00	-0.59	0.26	0.19
1.20	0.00	0.26	1.35	0.18	0.00	1.00	-0.31	-1.46	0.21
1.20	0.00	1.52	-0.03	0.16	0.00	0.70	-0.52	0.40	0.20
0.00	0.70	-1.25	-0.24	0.16	1.30	0.00	-0.10	0.74	0.18
0.00	0.90	-0.03	-1.04	0.20	0.00	0.80	-0.17	-0.38	0.22
1.60	0.00	0.53	-2.00	0.21	1.60	0.50	-0.38	-0.17	0.16
0.00	0.60	-0.38	-1.46	0.20	0.00	0.60	1.02	-2.00	0.21
2.00	0.00	0.68	0.62	0.22	2.00	0.00	-1.25	-0.84	0.20

CONFIGURATION 3 TWO TRAITS
BASE ITEM PARAMETERS

EASY TEST					HARD TEST				
AL1	AL2	BL1	BL2	CL	AH1	AH2	BH1	BH2	CH
0.50	0.60	0.32	-0.10	0.16	0.50	1.60	0.32	0.53	0.16
0.60	0.50	0.04	-0.45	0.21	0.60	1.30	-0.84	-0.84	0.16
0.60	0.60	-0.10	0.11	0.19	0.60	1.40	-0.66	-1.04	0.20
0.70	0.70	-0.59	0.32	0.20	0.70	2.00	-0.52	0.26	0.18
0.70	0.70	0.74	0.62	0.19	0.70	1.80	0.11	0.62	0.16
0.80	0.00	0.53	-2.00	0.21	0.80	0.00	-0.59	-0.66	0.17
0.80	0.00	0.68	-1.46	0.17	0.80	0.00	0.62	-0.45	0.21
0.90	0.00	-0.38	-0.03	0.16	0.90	0.00	0.26	1.02	0.20
0.80	0.00	0.62	1.35	0.16	0.80	0.00	1.35	-0.17	0.17
1.00	0.00	-0.66	-0.84	0.16	0.90	0.00	0.40	-0.52	0.16
0.90	0.00	0.95	-0.59	0.20	0.90	0.00	1.52	-1.46	0.21
0.90	0.00	1.02	0.40	0.18	0.80	0.00	0.04	0.68	0.16
0.80	0.00	0.49	-1.04	0.17	0.90	0.00	0.18	0.74	0.20
0.90	0.00	0.40	-1.25	0.20	1.00	0.00	-0.24	1.35	0.16
1.00	0.00	-1.46	-0.66	0.20	1.00	0.00	-0.10	0.11	0.18
1.00	0.00	-0.31	1.02	0.22	1.00	0.00	-0.38	-0.31	0.22
1.00	0.00	-0.24	0.95	0.16	1.00	0.00	1.02	0.49	0.22
1.20	0.00	-0.17	0.26	0.16	1.00	0.00	0.74	0.95	0.20
1.10	0.00	1.35	0.74	0.17	1.10	0.00	0.95	0.32	0.22
1.10	0.00	-2.00	0.68	0.16	1.10	0.00	-0.03	-1.25	0.17
1.10	0.00	-0.45	-0.31	0.20	1.10	0.00	-0.45	-0.59	0.19
1.10	0.00	-0.03	0.49	0.18	1.10	0.00	0.49	-0.38	0.18
1.00	0.00	0.26	0.53	0.22	1.20	0.00	0.53	0.18	0.20
1.20	0.00	1.52	0.04	0.22	1.80	0.00	-0.31	-0.03	0.20
1.20	0.00	0.18	1.52	0.19	1.20	0.00	0.68	-0.10	0.21
1.30	0.00	-0.84	0.18	0.21	1.30	0.00	-2.00	-0.24	0.16
1.40	0.00	-1.04	-0.24	0.20	1.40	0.00	-1.25	0.04	0.20
1.60	0.00	-1.25	-0.38	0.18	1.60	0.00	-0.17	-2.00	0.19
1.80	0.00	0.11	-0.17	0.20	1.20	0.00	-1.46	1.52	0.19
2.00	0.00	-0.52	-0.52	0.20	2.00	0.00	-1.04	0.40	0.20

TABLE 3 (CONT.)

CONFIGURATION 4 TWO TRAITS
BASE ITEM PARAMETERS

EASY TEST					HARD TEST				
AL1	AL2	BL1	BL2	CL	AH1	AH2	BH1	BH2	CH
0.50	0.00	-2.00	-0.38	0.18	0.90	0.00	-0.45	-0.10	0.21
0.90	0.00	-0.24	-0.66	0.20	0.90	0.00	-1.04	-2.00	0.16
0.60	0.00	0.40	-0.10	0.20	1.10	0.00	-0.66	-1.25	0.17
0.70	0.00	-0.59	-0.31	0.20	0.60	0.00	-0.52	-0.17	0.22
0.70	0.00	-1.46	1.35	0.21	0.70	0.00	0.04	-0.84	0.22
0.80	0.00	-1.04	-0.52	0.19	1.00	0.00	-0.24	-0.66	0.20
0.80	0.00	0.53	-0.84	0.20	0.80	0.00	-0.31	-0.59	0.17
0.80	0.00	-0.84	-0.45	0.17	0.80	0.00	-1.46	-1.04	0.20
1.20	0.00	-0.17	-0.03	0.22	2.00	0.00	0.40	-0.31	0.20
0.90	0.00	-1.25	0.32	0.16	1.00	0.00	1.52	0.74	0.16
0.90	0.00	0.68	0.04	0.20	1.20	0.00	0.68	-0.45	0.22
0.60	0.00	-0.31	-0.24	0.16	1.10	0.00	0.18	-0.24	0.18
0.90	0.00	0.26	-0.59	0.19	0.90	0.80	-0.84	1.02	0.16
1.00	0.00	-0.52	-1.25	0.21	1.30	0.00	-0.10	-0.52	0.16
1.00	0.00	0.62	0.11	0.22	1.40	0.00	0.74	0.04	0.20
1.00	0.00	0.95	0.26	0.18	0.80	0.50	-0.59	0.62	0.16
1.00	0.00	1.52	0.62	0.17	1.80	0.70	-0.17	0.95	0.17
1.10	0.00	-0.03	0.53	0.16	1.20	0.00	1.02	0.18	0.16
1.10	0.00	-0.45	0.49	0.16	0.70	0.00	-2.00	-1.46	0.18
1.00	0.00	0.49	0.68	0.17	1.10	0.00	0.11	0.32	0.20
1.10	0.00	-0.66	0.40	0.16	1.00	0.00	0.32	0.40	0.18
1.10	0.00	0.11	-1.46	0.20	0.60	0.00	-0.38	-0.03	0.20
0.80	0.00	0.74	-1.04	0.16	0.80	0.00	0.53	0.11	0.20
1.20	0.00	1.02	-2.00	0.21	0.90	0.00	0.95	0.49	0.19
1.20	0.00	-0.10	0.95	0.22	1.60	0.00	0.26	0.53	0.20
1.30	0.00	0.04	1.52	0.18	0.50	0.00	-1.25	-0.38	0.21
1.40	0.00	0.32	0.18	0.20	1.10	0.00	1.35	0.26	0.19
1.60	0.00	0.18	-0.17	0.16	1.20	0.00	-0.03	0.68	0.16
1.80	0.00	-0.38	0.74	0.20	1.00	1.00	0.49	1.52	0.19
2.00	0.00	1.35	1.02	0.19	1.00	0.90	0.62	1.35	0.21

CONFIGURATION 5 UNIDIMENSIONAL
BASE ITEM PARAMETERS

EASY TEST			HARD TEST		
AL1	BL1	CL	AH1	BH1	CH
0.80	0.68	0.18	0.90	0.62	0.16
2.00	0.18	0.20	0.90	-0.03	0.18
1.00	0.62	0.18	2.00	-0.31	0.17
1.10	-0.17	0.20	1.00	1.35	0.20
0.50	0.32	0.16	0.60	1.02	0.19
1.20	0.40	0.20	1.00	0.95	0.20
1.00	0.11	0.16	0.80	-0.84	0.16
1.20	-0.66	0.16	1.00	-2.00	0.18
1.20	-0.31	0.18	1.10	0.26	0.20
0.90	-0.84	0.20	0.60	-0.10	0.22
1.10	-2.00	0.19	1.10	0.32	0.22
0.60	-0.45	0.19	1.00	-0.45	0.20
0.80	0.49	0.16	0.50	0.49	0.21
1.00	-0.59	0.22	0.70	0.68	0.16
0.90	-0.52	0.17	1.00	-1.46	0.17
0.80	-1.25	0.20	0.80	-0.59	0.19
0.90	-0.24	0.20	1.20	-0.52	0.16
0.70	1.52	0.20	1.60	0.53	0.20
0.70	-0.10	0.16	0.90	-0.17	0.19
1.10	-0.38	0.16	1.80	0.04	0.17
0.60	-1.04	0.21	0.70	-1.04	0.16
1.60	0.95	0.21	1.30	-1.25	0.18
1.80	-1.46	0.20	0.80	-0.24	0.16
1.00	1.02	0.22	1.20	0.18	0.20
0.90	0.53	0.22	0.90	0.74	0.21
1.30	1.35	0.21	1.10	-0.66	0.16
1.40	0.04	0.16	0.80	-0.38	0.20
0.80	-0.03	0.17	1.10	0.11	0.21
1.00	0.26	0.17	1.40	0.40	0.20
1.10	0.74	0.19	1.20	1.52	0.22

TABLE 4

LOGIST Runs for Parameter Estimation

LOGIST Run	Sample	N	ITEMS Test 1 (30 Items)	Test 2 (30 Items)
1	Low Ability	2,000	Responses	NR
	Medium Ability	2,000	15 NR 15 Responses	15 NR 15 Responses
2	Medium Ability	2,000	15 NR 15 Responses	15 NR 15 Responses
	High Ability	2,000	NR	Responses

Note: "NR" means "not reached" was used in place of item responses.
"Responses" means the item responses were used.

TABLE 5

LOGIST Runs for Equating

LOGIST Run	Sample	N	Items Test 1	Items Test 2
1	All three ability levels.	6,000	Responses	NR
2	All three ability levels.	6,000	NR	Responses

Table 6

True Item Parameter Correlations Across and Within Traits

Configuration	Test	Desired				Attained			
		$r_{a_1a_2}$	$r_{a_1b_1}$	$r_{a_2b_2}$	$r_{b_1b_2}$	$r_{a_1a_2}$	$r_{a_1b_1}$	$r_{a_2b_2}$	$r_{b_1b_2}$
1	Test 1	0	0	0	0	.04	-.02	.05	.03
	Test 2	0	0	0	0	-.05	-.01	-.05	.05
2	Test 1	-.8	0	0	0	-.72	.01	-.09	.04
	Test 2	-.8	0	0	0	-.70	.07	-.01	.07
3	Test 1	-.5	-.3	0	0	-.54	-.31	.09	-.01
	Test 2	-.5	-.3	0	0	-.54	-.30	.01	-.06
4	Test 1	0	.4	0	0	.00	.39	.00	-.05
	Test 2	0	.4	.7	.5	.08	.40	.64	.54
			r_{ab}				r_{ab}		
U	Test 1		0				-.02		
	Test 2		0				-.01		

TABLE 7

Correlations Between Traits

for Generated True Thetas

	$r_{\theta_1\theta_2}$	
Configuration	Desired	Attained Range
1	.60	.59 to .63
2	.50	.48 to .52
3	.50	.47 to .53
4	.50	.47 to .51

TABLE 8

Means and Standard Deviations

for Generated True Thetas

	$\bar{\theta}$		SD_θ	
Ability	Desired	Attained Range	Desired	Attained Range
Low	-.57	-.67 to -.51	1.0	.96 to 1.04
Medium	0.00	-.04 to .03	1.0	.96 to 1.04
High	0.57	.54 to .61	1.0	.98 to 1.04

TABLE 9

Means, Standard Deviations, and KR-20 Values
of Number-Correct Scores for the Simulated Tests

Configuration	$\bar{b}_2-\bar{b}_1$	Test	Low + medium ability			Medium + high ability		
			$\bar{X}$	SD_x	KR-20	$\bar{X}$	SD_x	KR-20
1	2	1	23.20	7.08	.92	*	*	*
	1		19.71	8.03	.92	23.87	6.80	.92
	0		15.79	8.00	.91	20.33	7.89	.92
	0	2	15.99	7.76	.90	20.42	7.48	.91
	1		12.40	6.94	.87	16.44	7.75	.90
	2		9.63	5.69	.82	12.93	7.15	.88
2	2	1	22.69	5.53	.85	25.46	4.40	.83
	1		19.70	6.05	.85	23.01	5.46	.85
	0		16.84	5.98	.83	20.28	5.94	.85
	0	2	15.93	6.02	.83	19.41	6.01	.85
	1		12.99	5.54	.79	16.32	6.20	.84
	2		10.62	4.83	.74	13.35	5.61	.80
3	2	1	22.10	5.94	.87	24.94	4.87	.86
	1		19.32	6.45	.87	22.54	5.72	.86
	0		16.47	6.49	.86	19.97	6.36	.87
	0	2	16.25	6.36	.85	19.76	6.36	.87
	1		13.30	5.84	.82	16.54	6.42	.86
	2		10.60	5.02	.76	13.58	5.99	.83
4	2	1	21.94	6.04	.87	24.69	5.15	.86
	1		19.04	6.49	.87	22.17	6.04	.87
	0		16.15	6.52	.85	19.40	6.44	.87
	0	2	15.33	6.43	.85	18.72	6.62	.87
	1		12.69	5.66	.81	15.77	6.51	.86
	2		10.55	4.85	.75	12.98	5.82	.82
U	2	1	21.90	5.99	.86	24.60	5.05	.85
	1		19.04	6.54	.87	22.16	5.89	.86
	0		16.31	6.33	.85	19.59	6.35	.86
	0	2	16.45	6.37	.85	19.74	6.32	.86
	1		13.66	6.00	.82	16.65	6.38	.85
	2		11.37	5.21	.77	14.03	6.04	.83

* Incomplete LOGIST run due to too many perfect scores.

TABLE 10

Mean Item Chi-Squares (Q_1)

Configuration	$\bar{b}_2-\bar{b}_1$	Test	Ability Level	
			Medium+Low	Medium + High
1	2	1	29.9	*
	1		20.2	38.7
	0		32.5	22.9
	0	2	17.7	13.1
	1		19.6	11.0
	2		14.7	26.6
2	2	1	13.7	23.9
	1		14.3	19.6
	0		15.9	19.3
	0	2	17.6	13.5
	1		23.8	15.3
	2		20.2	16.4
3	2	1	10.8	24.6
	1		11.9	17.5
	0		13.8	18.3
	0	2	28.0	10.7
	1		35.9	12.6
	2		26.0	13.4
4	2	1	11.2	24.4
	1		9.6	14.3
	0		11.8	16.3
	0	2	21.2	10.0
	1		18.3	12.4
	2		22.9	12.8
U	2	1	10.6	21.3
	1		11.4	13.1
	0		13.8	12.6
	0	2	15.8	6.6
	1		16.9	21.1
	2		24.4	11.8

* Incomplete LOGIST run due to too many perfect scores.

TABLE 11

Mean Q_2 and Percent Q_2 Greater than 23.2

Configuration	$\bar{b}_2-\bar{b}_1$	Test	$\bar{Q}_2$		Percent $Q_2 > 23.2$	
			Low+Med Ability	Med+High Ability	Low+Med Ability	Med+High Ability
1	2	1	10.1	*	.04	*
	1		11.0	9.1	.05	.04
	0		11.8	10.8	.06	.05
	0	2	11.0	11.3	.05	.06
	1		11.9	12.0	.06	.07
	2		10.4	11.8	.05	.06
2	2	1	23.8	19.0	.40	.27
	1		25.8	24.4	.47	.42
	0		23.4	25.2	.42	.43
	0	2	23.5	23.6	.40	.40
	1		22.8	25.4	.33	.42
	2		18.3	21.2	.20	.31
3	2	1	11.9	11.0	.08	.07
	1		12.8	11.4	.08	.06
	0		13.4	12.5	.09	.08
	0	2	15.8	15.2	.14	.12
	1		15.2	15.8	.12	.14
	2		13.9	15.0	.11	.12
4	2	1	11.8	11.3	.06	.06
	1		13.0	12.0	.05	.07
	0		13.2	12.3	.10	.04
	0	2	13.8	12.9	.09	.08
	1		13.6	13.2	.09	.09
	2		14.5	14.0	.10	.10
U	2	1	11.6	10.5	.07	.05
	1		12.5	11.0	.06	.06
	0		12.8	12.0	.06	.05
	0	2	12.6	11.9	.06	.06
	1		12.6	12.8	.08	.05
	2		14.2	13.6	.08	.08

* Incomplete LOGIST run due to too many perfect scores.

TABLE 12

Mean, Standard Deviation, and Correlation of Estimated Thetas for Test 1 versus Test 2

Configuration	$\bar{b}_2 - \bar{b}_1$	Test 1 $\bar{\hat{\theta}}_1$	Test 1 SD_1	Test 2 $\bar{\hat{\theta}}_2$	Test 2 SD_2	$r_{\hat{\theta}_1\hat{\theta}_2}$
1	0	.21	.72	.15	.68	.89
	1	.21	.70	.32	.50	.84
	2	*	*	.63	.45	*
2	0	.28	1.14	.21	1.23	.83
	1	.26	.94	.32	.92	.83
	2	.18	.78	.30	.76	.72
3	0	.25	.98	.26	1.07	.85
	1	.26	.89	.29	.85	.83
	2	.26	.75	.38	.77	.75
4	0	.20	1.05	.22	1.19	.84
	1	.21	.99	.22	1.14	.81
	2	.18	.82	.26	1.09	.70
U	0	.25	1.01	.25	.95	.86
	1	.26	.88	.31	.73	.83
	2	.18	.85	.20	.96	.74

* Incomplete LOGIST runs due to too many perfect scores on Test 1.

Note: Test 1 is the easier test and Test 2 is the harder.

TABLE 13

Standardized Root Mean Squared Differences, Standardized Differences between Means, Ratios of Standard Deviations, and Local Bias for Test 1 versus Test 2

Configuration	$\bar{b}_2-\bar{b}_1$	SRMSD	SD_1/SD_2	SDM	Local Bias
1	0	.49	1.07	.07	.07
	1	.60	1.18	-.15	.15
	2	*	*	*	*
2	0	.59	.96	.06	.06
	1	.59	1.01	-.07	.06
	2	.76	1.01	-.13	.13
3	0	.54	.96	-.01	.04
	1	.59	1.02	-.03	.03
	2	.72	.99	-.14	.13
4	0	.57	.94	-.02	.04
	1	.62	.93	-.01	.06
	2	.79	.87	-.08	.18
U	0	.54	1.03	-.01	.04
	1	.60	1.10	-.05	.09
	2	.73	.94	-.02	.03

Note: A positive SDM indicates that the Test 1 Mean is higher than the Test 2 Mean.

* Incomplete LOGIST run due to too many perfect scores on Test 1.

TABLE 14

Local Equating Bias for Estimated Thetas for Test 1 versus Test 2

Configuration	$\bar{b}_2-\bar{b}_1$	N**	Quintile				
			1	2	3	4	5
1	0	4742	.03	.09	.02	.03	.16
	1	4167	-.51	-.08	-.05	-.05	-.05
	2	*	*	*	*	*	*
2	0	5665	.13	.11	.06	.01	-.00
	1	5381	-.04	-.11	-.10	-.07	-.00
	2	4832	-.13	-.16	-.17	-.08	-.11
3	0	5469	.07	.03	-.02	-.04	-.06
	1	5189	-.08	.01	-.04	-.03	-.01
	2	4752	-.20	-.09	-.10	-.08	-.21
4	0	5629	.04	-.00	-.02	-.04	-.10
	1	5440	.12	-.05	-.03	.01	-.10
	2	4895	.24	-.17	-.19	-.12	-.16
U	0	5492	-.10	.04	.03	.01	-.00
	1	5223	-.17	-.12	-.06	-.01	.09
	2	5023	.04	-.07	-.03	-.02	-.00

Note: Quintiles range from 1 for the lowest ability to 5 for the highest. A positive difference indicates that the Test 1 mean is higher than the Test 2 mean.

* Incomplete data due to too many perfect scores in Test 1.

** Number of simulees for whom thetas were estimated on both tests.

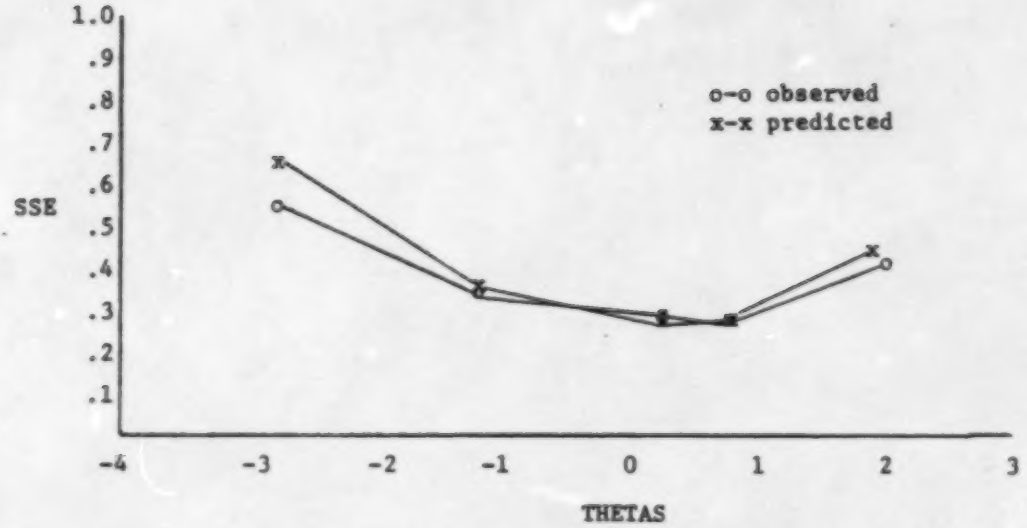

Figure 1. Observed and model predictions of standardized standard errors by quintile for Configuration U, $\bar{b}_2 - \bar{b}_1 = 0$.
Observed SSE = SRMSD/$\sqrt{2}$; predicted SSE = $\frac{\sqrt{(SE_1^2 + SE_2^2)/2}}{\sqrt{(S_1^2 + S_2^2)/2}}$.

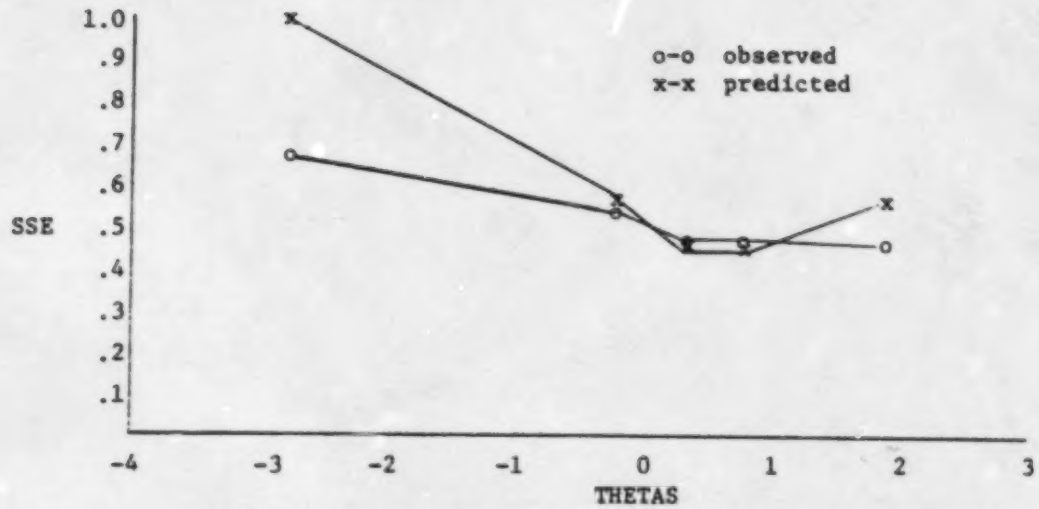

Figure 2. Observed and predicted standardized standard errors by quintile for Configuration 2, $\bar{b}_2 - \bar{b}_1 = 2$.

Observed SSE = SRMSD/$\sqrt{2}$; predicted SSE = $\dfrac{\sqrt{(SE_1^2 + SE_2^2)/2}}{\sqrt{(S_1^2 + S_2^2)/2}}$.